Nearly neutral evolution across the Drosophila melanogaster genome

Article (Accepted Version)

This is a pre-copyedited, author-produced version of an article accepted for publication in [insert journal title] following peer review. The version of record Castellano et al. (2018) Mol. Biol. Evol. 35, 2685-2694 is available online at: https://academic.oup.com/mbe/article/35/11/2685/5078937?searchresult=1.
doi:10.1093/molbev/msy164

Nearly Neutral Evolution Across the *Drosophila melanogaster* Genome

David Castellano[1]
Jennifer James[2]
Adam Eyre-Walker[2,3]

1. Bioinformatics Research Centre
Aarhus University
C.F. Møllers Allé 8
DK-8000 Aarhus C
Denmark

2. School of Life Sciences
University of Sussex
Brighton
BN1 9QG
United Kingdom

3. Correspondence: a.c.eyre-walker@sussex.ac.uk

Abstract

Under the nearly neutral theory of molecular evolution the proportion of effectively neutral mutations is expected to depend upon the effective population size (N_e). Here we investigate whether this is the case across the genome of *Drosophila melanogaster* using polymorphism data from North American and African lines. We show that the ratio of the number of non-

synonymous and synonymous polymorphisms is negatively correlated to the number of synonymous polymorphisms, even when the non-independence is accounted for. The relationship is such that the proportion of effectively neutral non-synonymous mutations increases by ~45% as N_e is halved. However, we also show that this relationship is steeper than expected from an independent estimate of the distribution of fitness effects from the site frequency spectrum. We investigate a number of potential explanations for this and show, using simulation, that this is consistent with a model of genetic hitch-hiking: genetic hitch-hiking depresses diversity at neutral and weakly selected sites, but has little effect on the diversity of strongly selected sites.

Introduction

In 2015, Tomoko Ohta was awarded the Crafoord prize, along with Richard Lewontin, for her contributions to evolutionary biology and population genetics. Among her many contributions, she is best known for her nearly neutral theory of molecular evolution in which she proposed that there is a class of mutations that are substantially influenced by both random genetic drift and natural selection (Ohta and Kimura 1971; Ohta 1972b, 1973, 1977, 1992). In species with large effective population sizes selection is effective and in species with small effective population sizes random drift dominates the fate of this class of mutations. As a consequence, the proportion of mutations that are effectively neutral varies with effective population size. The theory has had a profound influence on how we think about molecular evolution and processes such as the rate of evolution (Lanfear, et al. 2014), the evolution of the mutation rate (Lynch 2010; Lynch, et al. 2016) and genome size (Lynch and Conery 2003).

There are a number of lines of evidence suggesting that there is a class of mutations that are nearly neutral, the vast majority of which are thought to be slightly deleterious. First, putatively functional mutations segregate at lower frequencies in the population than putatively neutral mutations; for example, in many species non-synonymous mutations segregate at lower frequencies than synonymous mutations (Akashi 1999; Cargill, et al. 1999; Hughes 2005),

and it has also been shown that mutations in conserved non-coding sequences segregate at lower frequencies than those in neighbouring sequences (Andolfatto 2005; Drake, et al. 2006; Asthana, et al. 2007). Second, the rate of non-synonymous to synonymous substitution is higher along lineages leading to species with small effective population size (note that in most cases the effective population is not directly measured but a surrogate measure, such as body size, is used instead) (Ohta 1993, 1995; Moran 1996; Woolfit and Bromham 2003, 2005; Popadin, et al. 2007). In fact, Ohta produced the very first evidence for her theory by noting that the rate of divergence in protein coding sequences relative to the overall rate of DNA divergence, as measured by DNA-DNA hybridisation, was positively correlated to generation time, a proxy for population size (Ohta 1972a). Third, the ratio of non-synonymous or synonymous polymorphism appears to be greater in those species with smaller effective population size, as measured by synonymous diversity (Piganeau and Eyre-Walker 2009; Elyashiv, et al. 2010; Galtier 2016; Chen, et al. 2017; James, et al. 2017).

Ohta used her theory to explain how the rate of protein evolution could be relatively constant even if the mutation rate varied between taxa; she reasoned that species with large population sizes might have high mutation rates per year, but this would be offset by having a small proportion of effectively neutral mutations (Ohta and Kimura 1971; Ohta 1972b, 1992). However, despite all the qualitative evidence that nearly neutral mutations exist, there are few robust estimates how rapidly the proportion of effectively neutral mutations changes with effective population size.

We can potentially investigate how the proportion of effectively neutral mutations changes as a function of N_e by considering how the number of non-synonymous (p_N) relative to synonymous polymorphisms (p_S) changes as a function of the number of synonymous polymorphisms (p_S). The logic is as follows. Let us assume that synonymous mutations are neutral so that $p_S = kN_eu$ where k is Watterson's coefficient (=Sum $1/i$), N_e is the effective population size and u is the mutation rate. Let us further assume that non-synonymous mutations are either effectively neutral or deleterious so that $p_N =$

$kN_e uf$, where f is the proportion of effectively neutral mutations and k is Watterson's constant (Watterson 1975). We can therefore investigate how rapidly the proportion of effectively neutral mutations changes as N_e increases by considering the correlation between p_N/p_S and p_S. However, there are two issues. First, the two variables are not independent. We circumvent this problem by dividing p_S into two halves using a hypergeometric distribution (Piganeau and Eyre-Walker 2009; James, et al. 2017); this means that p_{S1} and p_{S2} have uncorrelated sampling errors. We then consider the correlation between p_N/p_{S1} and p_{S2}. Second, the theory outlined above assumes that the population is at equilibrium, however, non-equilibrium processes can potentially mimic the effects of variation in N_e. Brandvain and Wright (Brandvain and Wright 2016) give an instructive example. Imagine a population, initially with no diversity. The population will accumulate genetic diversity through mutation, but the sites subject to deleterious mutation will approach their equilibrium frequency faster than neutral mutations because the equilibrium frequency is lower for these sites (Gordo and Dionisio 2005; Do, et al. 2015; Brandvain and Wright 2016). Hence, if we were to sample a locus through time in this population, we would observe a negative relationship between p_N/p_S and p_S, despite the fact that N_e has not changed since the population was founded.

The precise relationship between p_N/p_S and p_S is potentially predictable. Ohta showed that the rate of evolution is proportional to $1/N_e$ over a substantial range of N_e values if the DFE is an exponential distribution (Ohta 1977). Kimura subsequently showed that under a gamma distribution with a shape parameter of 1/2 the rate was proportional to $1/ N_e^{0.5}$, which suggests, given that the exponential distribution is a gamma distribution with shape parameter of one, that the rate of evolution under a gamma distribution is proportional to $1/N_e^{\beta}$, where β is the shape parameter. This general result was confirmed by Welch et al.(Welch, et al. 2008). Welch et al. (2008) also demonstrated that the number of polymorphisms and the nucleotide diversity are proportional to $N_e^{1-\beta}$; hence the nucleotide diversity at selected sites divided by the nucleotide diversity at neutral sites should be proportional to $N_e^{-\beta}$ and hence

there should be a linear relationship, with a slope of $-\beta$ between the log of p_N/p_S and the log of p_S if the DFE is a gamma distribution.

Several groups have shown that log(p_N/p_S) is negatively correlated to log(p_S) across animal and plant species for nuclear DNA (Galtier 2016; Chen, et al. 2017) and across animal species for mtDNA (James, et al. 2017). It has also been demonstrated that this relationship holds across a single genome in *Capsella grandiflora, Arabidopsis lyrate* (Gossmann, et al. 2011) and the passenger pigeon (Murray, et al. 2017). This might be expected since the effective population size also varies across the genome in many species due to the effects of recombination, genetic hitch-hiking and background selection (Charlesworth 2009; Ellegren and Galtier 2016); regions of the genome with low rates of recombination, high mutations rates, gene densities or both, have lower effective population sizes than regions with high rates of recombination and a low density of selected mutations. However, no detailed study has been made within a species, and in particular, whether the relationship between log(p_N/p_S) and log(p_S) is quantitively consistent with the nearly neutral theory of molecular evolution.

Results

To investigate the relationship between the proportion of effectively neutral mutations and the effective population size, we compiled polymorphism data from 7918 autosomal genes from a North American population of *D. melanogaster*. We also use a smaller data set of 4676 autosomal genes with short introns (< 66 bp) as an alternative neutral reference (Halligan and Keightley 2006). Because we are interested in the relationship between log(p_N/p_S) and log(p_S) and p_N and p_S can be zero for individual genes, we grouped genes into 10 groups of 791 genes (10 groups of 467 genes when using introns) and considered the correlation between log(sum(p_N)/sum(p_{S1})) and log(sum(p_{S2})) (similar results were obtained with other groupings – see below).

As expected, given previous work, there is a positive correlation between p_S and the rate of recombination (RR) (Spearman's correlation $r_s = 0.99$, $p<0.001$) (Figure 1)(Begun and Aquadro 1992; Presgraves 2005; Langley, et al. 2012; Mackay, et al. 2012; Campos, et al. 2014). This positive correlation might be due to variation in the mutation rate or variation in N_e across the genome. Since, there is no correlation between the synonymous site divergence (d_S) between *D. melanogaster* and *D. yakuba* and RR ($r_s = -0.10$, $p = 0.78$), we conclude as others have done, that recombination is not mutagenic in Drosophila (Begun and Aquadro 1992; Betancourt and Presgraves 2002; Mackay, et al. 2012; Campos, et al. 2014), and that there is variation in N_e due to either genetic hitch-hiking or background selection. The relationship between p_S and RR is non-linear with a tendency for the number of synonymous polymorphisms to be substantially depressed in regions of low recombination (Figure 1).

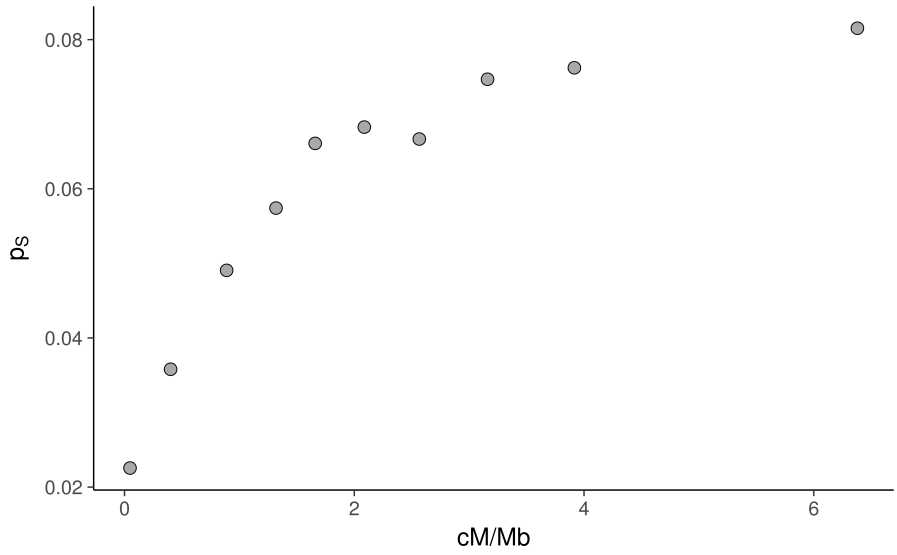

Figure 1. The correlation between the number of synonymous polymorphisms per site and the rate of recombination. Each point represents the average of 791 genes grouped according to their recombination rate.

We find that $\log(p_N/p_S)$ is negatively correlated to $\log(p_S)$ (slope = -0.53 (95% CIs = -0.48, -0.57), $p < 0$ .001) (figure 2A). The relationship is almost perfectly linear, consistent with the underlying distribution of fitness effects being a

gamma distribution. The slope is such that halving the effective population size increases the proportion of effectively neutral mutations by ~45%. The results are unaffected by using different grouping schemes (20 groups slope = -0.54 (-0.60, -0.48), 50 groups slope = -0.53 (-0.59, -0.48)).

We have assumed that synonymous mutations are neutral, whereas we know that there is selection on synonymous codon use in *Drosophila* species (Shields, et al. 1988; Akashi 1995). To investigate whether this affects our results we repeated the analysis using data from short introns as our neutral sites (Halligan and Keightley 2006). The relationship between $\log(p_N/p_{I1})$ and $\log(p_{I2})$ is highly significant ($p < 0.001$) (figure 2B), almost perfectly linear and has a slope which is very similar to that found using synonymous sites (slope = -0.55 (-0.47, -0.65)). However, if there is selection on short introns then this would affect our conclusions.

In many species, we do not have a recombination rate map, so we also investigated the relationship between $\log(p_N/p_S)$ and $\log(p_S)$ grouping genes by their p_S or p_I values rather than the RR. To do this we followed the method of James et al. (James, et al. 2017) in which p_S (p_I) is split three-ways using a hypergeometric distribution and p_{S1} (p_{I1}) is used to rank and group genes, p_{S2} (p_{I2}) is used to measure p_S (p_I) and p_{S3} (p_{I3}) is used to calculate p_N/p_S (p_N/p_I). The slopes using this method are -0.52 (-0.48, -0.56) and -0.64 (-0.53, -0.77) for synonymous and intron sites respectively (Figure 2C, 2D), confirming that this is a satisfactory method for investigating the relationship between $\log(p_N/p_S)$ and $\log(p_S)$ when recombination rate data are not available and there is sufficient polymorphism data, although it should be noted that the results are subject to higher levels of sampling error.

We performed most of our analyses using the DGRP lines which were sampled from a derived population in the United States (Mackay, et al. 2012). To check that our results are consistent across populations we also performed the analysis on the DPGP2 lines sampled from Rwanda (Pool et al. 2012). Whether we group genes by RR or p_S we observe a negative correlation

between log(p_N/p_S) and log(p_S) (Figure 2E, 2F) (grouping by *RR* slope = -0.42 (-0.34, -0.50); grouping by p_S slope = -0.51 (-0.43, -0.58).

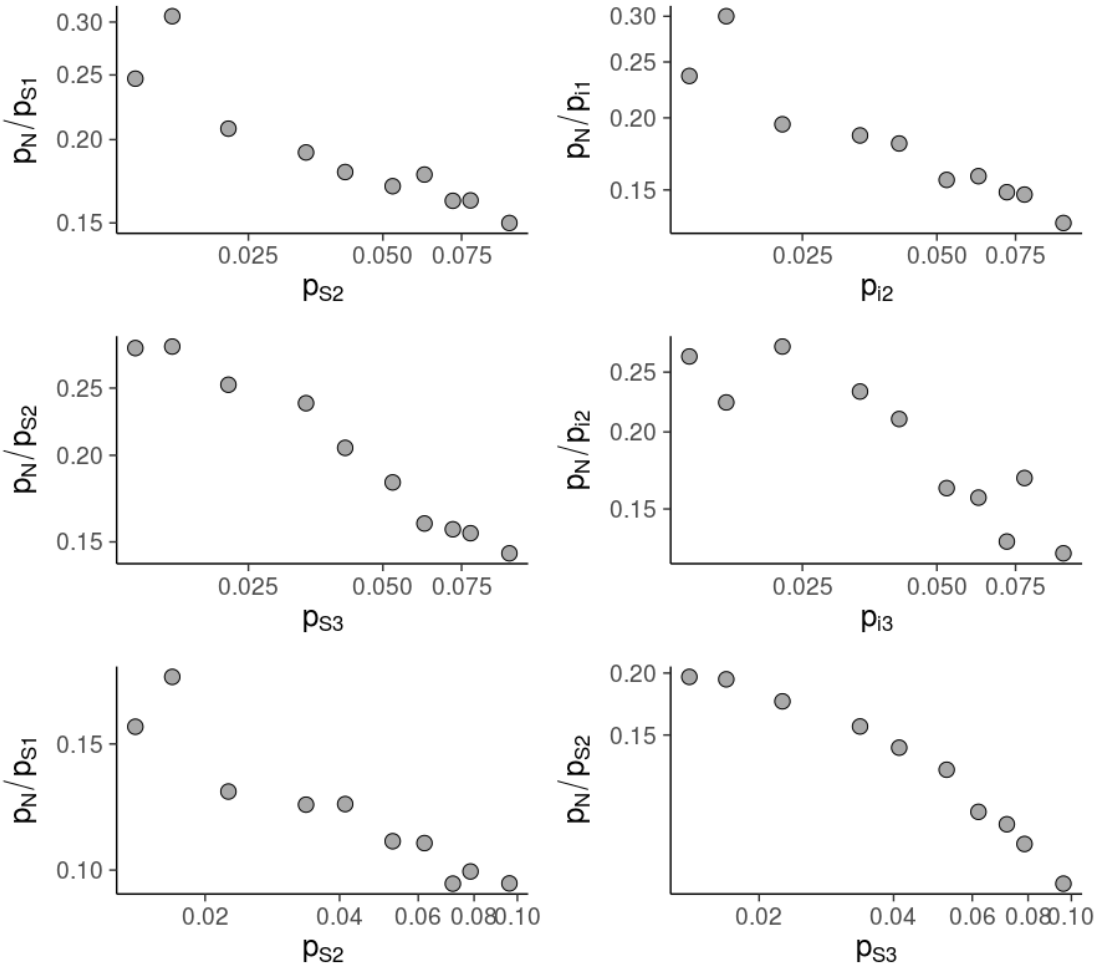

Figure 2. The relationship between p_N/p_S and p_S or p_i grouping genes by either recombination rate, p_S or p_i. Panels A to D are using DGRP flies, panels E and F are using DPGP flies. A) using p_s as the neutral standard, grouping genes by recombination rate, B) using p_i, grouping genes by recombination rate, C) using p_s, grouping genes by p_s, D) using p_i, grouping genes by p_i, E) using p_s, grouping genes by recombination rate and F) using p_s, grouping genes by p_s.

If the DFE is gamma distribution, then we expect the slope of the relationship between $\log(p_N/p_S)$ and $\log(p_S)$ to be equal to the negative of the shape parameter of the DFE (Welch, et al. 2008). To investigate whether this is the case we estimated the shape parameter of the DFE using the site frequency spectrum (SFS) at non-synonymous and synonymous sites using the method

of Keightley and Eyre-Walker (Keightley and Eyre-Walker 2007). We find that the shape parameter is significantly smaller (shape = 0.33 (0.30, 0.38)) than the slope of the relationship between log(p_N/p_S) and log(p_S) ($p < 0.001$ bootstrapping the data by gene) for the American DGRP data and the African DPGP2 data (shape parameter = 0.35 (0.34, 0.36) which is significantly shallower than the slope, $p<0.001$). This discrepancy has been noted before across species for animal mtDNA (James, et al. 2017) and although not explicitly mentioned by the authors, is evident in the analysis between species in animal and plant nuclear DNA (Chen, et al. 2017).

There are a number of potential explanations for why the slope of the relationship between log(p_N/p_S) and log(p_S) is greater than the shape parameter of the DFE estimated using the SFS. First, there could be problems with our method. However, previous simulations have suggested that the method recovers the expected relationship between log(p_N/p_S) and log(p_S) under a gamma DFE in a model with background selection and linkage (James, et al. 2017).

Second, the pattern may be a product of inversions, admixture or identity-by-descent (IBD). Both datasets are known to have some admixture (Pool, et al. 2012; Pool 2015) and low recombination regions in the DGRP data show elevated admixture (Pool 2015), so our results could be affected by admixture, given the relationship between p_S and RR. Similarly, inversion polymorphisms appear to increase levels of genetic diversity (Corbett-Detig and Hartl 2012; Pool, et al. 2012). We have removed inversions from the DGRP, and Campos et al. (Campos, et al. 2014) removed admixed regions from the DPGP data. In neither dataset were regions of IBD removed but these comprise a small fraction of the Rwandan dataset, in which this complication has been quantified – two individuals are estimated to be affected over 9 and 21MB of their genome (Pool, et al. 2012). It seems unlikely that inversions, admixture and IBD are responsible for our results given that the patterns we observe are consistent across two datasets which have different origins and have been processed differently.

Third, there might be a negative relationship between N_e and the mutation rate. This seems unlikely within the *Drosophila* genome, given that there is no correlation between d_S, a measure of the mutation rate, and the rate of recombination (see above). However, to investigate further, we split the synonymous (and intronic) divergence into two halves using the hypergeometric distribution then used d_{S1} (d_{I1}) to calculate the average N_e for a group of genes (grouped by RR) as p_S/d_{S1} (p_I/d_{I1}) and used the other d_S (d_I) value as our estimate of the mutation rate. For synonymous sites, we find no correlation between our estimate of N_e and d_S (slope = -0.017, p = 0.54) but for introns we observe a positive and significant relationship (slope = 0.10, p < 0.001) (Supplementary Figures 1A and 1B); this latter correlation is consistent with the fact that d_I is positively correlated to RR at the gene level (r_s=0.09, p<0.001). These patterns are consistent with there being a positive correlation between RR and the mutation rate, a positive relationship which is offset at synonymous sites by greater efficiency of selection against deleterious mutations in highly recombining areas, leading to a lack of a correlation between synonymous divergence and RR. It therefore seems likely that there is a positive correlation between the mutation rate and N_e in *Drosophila* and hence no evidence that they are negatively correlated. Similar results are obtained for the relationship between $\log(p_N/p_S)$ versus $\log(p_S/d_S)$ and $\log(p_N/p_i)$ versus $\log(p_i/d_i)$ grouping by RR (using synonymous SNPs, slope = -0.50 (-0.48, -0.57); using intron SNPs, slope = -0.60 (-0.51, -0.70) or neutral polymorphism levels (-0.55 (-0.50, -0.59) and -0.54 (-0.47, -0.62)).

Fourth, the slope of the relationship between $\log(p_N/p_S)$ and $\log(p_S)$ might be steeper than the estimate of the shape parameter of the DFE because the DFE is affected by N_e; the shape parameter or the mean selection coefficient of new deleterious mutations might increase with increasing N_e. To investigate this, we estimated the DFE using the SFS for each group of genes separately, grouping genes by the RR. The method estimates the shape parameter and the mean value of N_es for the DFE. Since we expect mean N_es to increase as N_e increases, we divided our estimate of mean N_es by p_S to yield an estimate of the mean strength of selection (equivalent results are obtained if N_es is divided by p_S/d_S). There is a marginally significant positive correlation

between p_S and the shape parameter (r = 0.72, p = 0.020) (Figure 3a), however the slope of this relationship is very shallow, and the shape parameter never approaches the (negative) of the slope of the relationship between log(p_N/p_S) and log(p_S). The mean strength of selection is uncorrelated to recombination rate (r = 0.51, p = 0.13)(Figure 3b), but this is not a powerful analysis since the estimate of the mean value of $N_e s$ is subject to considerable uncertainty, even in a large sample such as the one we have used.

Fifth, the difference between the slope of the relationship between log(p_N/p_S) and log(p_S) and the shape parameter of the DFE estimated from the SFS might simply be due to the fact that the DFE is not well described by the gamma distribution and hence there is no expectation that the two quantities should be similar. The relationship between log(p_N/p_S) and log(p_S) is very close to being linear for DGRP and DPGP data, suggesting that the gamma is a reasonable fit to the data. However, other distributions might give equally good fits.

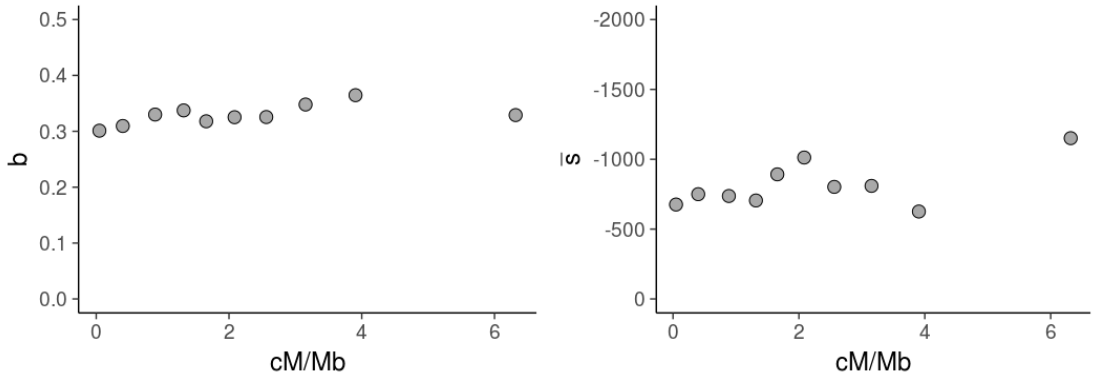

Figure 3. The relationship between the shape parameter (left panel) and the estimated mean strength of selection (right panel) and the rate of recombination.

Sixth, linked selection might affect the effective population size of neutral and deleterious mutations to different extents. We have previously shown, using simulation, that background selection does not affect the slope of the relationship between log(p_N/p_S) and log(p_S) (James, et al. 2017), however hitch-hiking may affect the slope. It has been demonstrated, using simulated data, that deleterious genetic diversity recovers more rapidly after a bottleneck than neutral genetic diversity (Gordo and Dionisio 2005; Brandvain and Wright 2016) and this is also evident in HIV data after a hard-selective sweep (Pennings, et al. 2014).

To investigate whether hitch-hiking is responsible for the increase in the slope, we ran two sets of simulations. In the first, we simulated a non-recombining locus composed of neutral sites and blocks of sites subject to varying levels of natural selection (i.e. the locus is composed of an equal number of sites subject to Ns = -1, -2, -4…etc). This locus was also subject to adaptive evolution, the rate of which was varied between simulations. We find, as expected, that nucleotide diversity for neutral and weakly selected sites decreases at the locus as we increase the rate of adaptive mutation (Figure 4a). However, the effect is attenuated for mildly deleterious mutations and there is no effect on diversity for strongly selected deleterious mutations,

under the conditions of our simulation. As expected, the effect is attenuated if there is recombination between the advantageous locus and locus with neutral and deleterious mutations (Figure 4b,c, d). In these simulations we used a population size of 100, but similar results were obtained with 500 individuals (Figure S1).

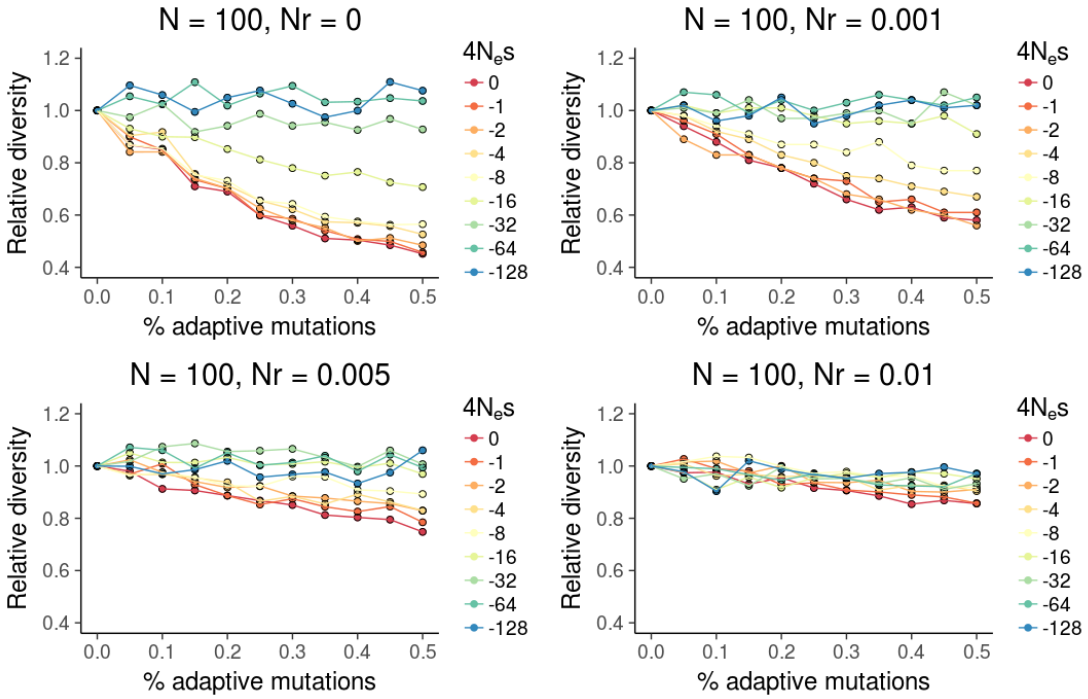

Figure 4. The effect of genetic hitch-hiking. The diversity at sites subject to different levels of negative selection, but all part of the same non-recombining locus, plotted against the frequency of advantageous mutation in the locus, for different population sizes and rates of recombination between the locus with advantageous mutations and that with deleterious and neutral mutations.

The fact that hitch-hiking depresses the diversity at neutral and weakly selected sites more than at strongly selected sites could potentially explain why the slope of $\log(p_N/p_S)$ and $\log(p_S)$ is steeper than expected from the estimate of the shape parameter of the DFE; regions with high levels of hitch-hiking will have low diversity, but the diversity of non-synonymous sites will be depressed less than that of synonymous sites; hence the slope between $\log(p_N/p_S)$ and $\log(p_S)$ will be steeper. To investigate this, we ran a second set of simulations in which a locus, that experienced both neutral and deleterious mutations, whose fitness effects were drawn from a gamma distribution occurred, were linked to a locus undergoing adaptive evolution. For each rate of adaptive evolution, we averaged the number of neutral and selected polymorphisms per site across multiple sampling points and runs. We find the slope of $\log(p_N/p_S)$ and $\log(p_S)$ is substantially and significantly steeper (slope

= -0.67 (0.029)) than -0.4 (p<0.001), the negative of the shape parameter of the gamma distribution used to generate the deleterious mutations (Figure 5). Similar results were obtained using a population size of 500 (Figure S1); in this case the slope = -0.74 which is significantly greater than the expected value of -0.4 (p<0.001).

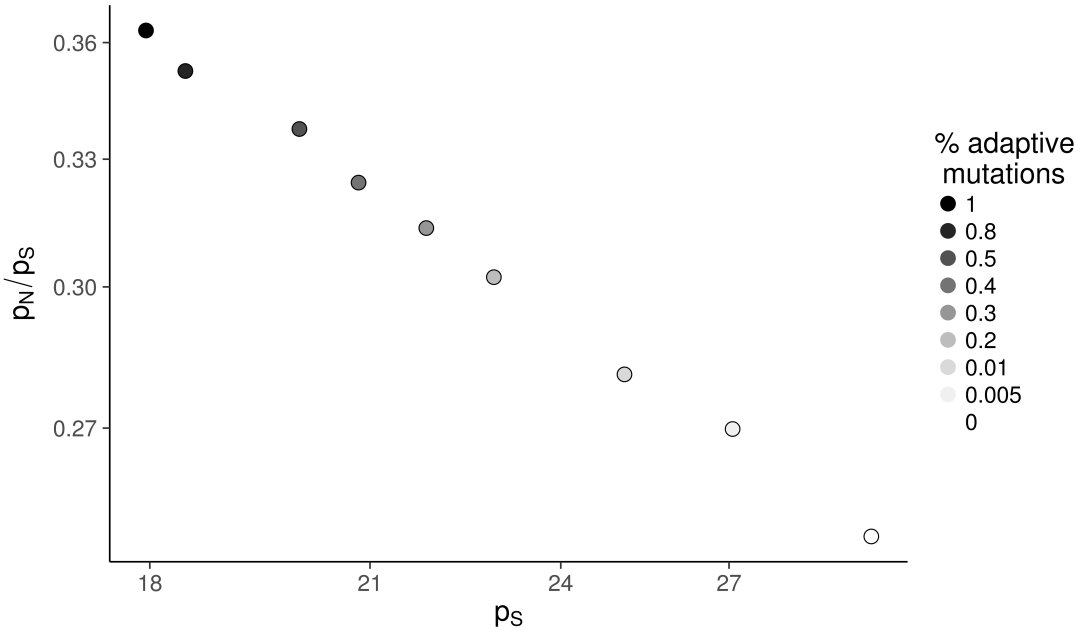

Figure 5. The effects of hitch-hiking. p_N/p_S plotted against p_S both on log-scales, for simulations in which deleterious mutations are drawn from a gamma distribution for different rates of adaptive evolution. Each point represents the mean from a simulation run at a certain rate of beneficial mutation. The slope of the line, -0.67, is significantly steeper than the shape parameter of the gamma distribution used to simulate the data, which was 0.4.

Discussion

We have shown that the proportion of mutations that are effectively neutral, as measured by the ratio of the number of non-synonymous and synonymous polymorphisms, decreases as the effective population size, as measured by the number of synonymous polymorphisms, increases across the *D. melanogaster* genome. On a log-log scale the relationship is approximately

linear and the slope of the relationship is such that halving the effective population size increases the proportion of effectively neutral mutations by ~45%. The slope of the relationship between log(p_N/p_S) and log(p_S) is similar to that found across plant and animal species for nuclear DNA (Galtier 2016; Chen, et al. 2017) and for animal mtDNA (James, et al. 2017).

Using data from across the *D. melanogaster* genome, we find, as others have between species (Chen, et al. 2017; James, et al. 2017), that the slope of the relationship between log(p_N/p_S) and log(p_S) is steeper than we expect given an estimate of the DFE from the site frequency spectrum. We have previously shown by simulation that our method recovers the expected relationship when non-synonymous mutations are gamma distributed, even if there is substantial linkage between sites, so it seems unlikely that the steeper slope is due to a problem with our method (James, et al. 2017). We also find no evidence that there is a negative correlation between N_e and the mutation rate across the genome or a change in the shape parameter or mean of the DFE. However, it is possible that the discrepancy is because the DFE is not gamma distributed. Although, the relationship between log(p_N/p_S) and log(p_S) is approximately linear, as expected under the gamma distribution (Ohta 1977; Welch, et al. 2008), it is possible that other distributions, such as the log-normal would fit the data equally well. It is also possible that our estimate of the DFE is being influenced by mutations that are positively selected, particularly those subject to balancing selection. The influence of positively selected mutations on the estimation of the DFE and the relationship between log(p_N/p_S) and log(p_S) is unknown.

We have shown that the steeper slope is consistent with a model of genetic hitch-hiking. We find, through simulation, that genetic hitch-hiking affects the relationship between log(p_N/p_S) and log(p_S), making it steeper than expected. We demonstrate that hitch-hiking depresses the level of genetic diversity at neutral and slightly deleterious sites, but has little or no effect on the diversity of strongly deleterious mutations. As a consequence, regions of the genome with high rates of hitch-hiking have low neutral diversity, but deleterious genetic diversity is not as affected. This effect is likely due to the fact that

genetic diversity recovers more rapidly at negatively selected sites than neutral sites after a bottleneck or hitch-hiking event (Gordo and Dionisio 2005; Pennings, et al. 2014; Brandvain and Wright 2016). The finding that hitch-hiking can induce a relationship between log(p_N/p_S) and log(p_S) raises the question of whether the relationship is due to variation in the efficiency of selection, non-equilibrium dynamics, or a combination of both. Resolving this question will require knowledge about the rates and strength of selection acting upon advantageous mutations and extensive simulation. The effect of hitch-hiking may also depend on demography – our simulation does not attempt to mimic the demographic changes that have affected *D. melanogaster.*

It is currently unclear how general these results are likely to be; are we likely to find a negative correlation between p_N/p_S and p_S across other genomes? This will depend upon variation in the rate of recombination and density of selected sites establishing variation in N_e, the distribution of fitness effects and/or variation in the rate of adaptive evolution across the genome. Gossman et al. (Gossmann, et al. 2011) and Murray et al. (Murray, et al. 2017) have recently shown that p_N/p_S is negatively correlated to p_S across the two plant species and the passenger pigeon genome. Also many species appear to show similar levels of variation in N_e across their genomes as *D. melanogaster* (Gossmann, et al. 2011) and hence might be expected to show a correlation between p_N/p_S and p_S; initial work suggests that this is the case in both primates and rodents, but with some qualitative differences (Castellano and Eyre-Walker, unpublished results).

An open question is whether variation in p_N/p_S across a genome has any implications for phenotypic variation. Whether it does, depends on the genetic architecture of quantitative traits; is most of the variation due to common or rare mutations of large or small effect, and what role does natural selection play in shaping this architecture? It seems likely that most of the variation in fitness will be contributed by rare mutations of relatively large effect if genetic variation is maintained in mutation-selection balance (Eyre-Walker 2010). As a consequence, variation in p_N/p_S is unlikely to be important because neither

variation in N_e or non-equilibrium dynamics will affect the equilibrium frequencies of strongly deleterious mutations. Variation in N_e, generated by for example background selection, has no effect, because the equilibrium frequencies of a deleterious mutation are independent of N_e, and we have shown that hitch-hiking also has little effect on the diversity of strongly selected mutations. However, weakly selected mutations may contribute substantially to traits that are not subject to strong selection and hence variation in p_N/p_S across the genome may affect the architecture of these traits.

Our results suggest that the proportion of effectively neutral mutations varies across the *Drosophila* genome and declines as a function of synonymous diversity; this is likely to be due to two factors; variation in the effective population size and non-equilibrium dynamics. Interestingly, we show that genetic hitch-hiking has little or no effect on strongly deleterious genetic variation.

Materials and methods

Datasets

This study was carried out on the four large autosomes (2L, 2R, 3L and 3R) of *D. melanogaster*. The population genomic data came from two sources, Raleigh, North Carolina (DGRP) (Mackay, et al. 2012) and Gikongoro, Rwanda (DPGP)(Pool, et al. 2012). We used Freeze 1.0 *Drosophila melanogaster* Genetic Reference Panel (DGRP) project. Sites with residual heterozygosity and low-quality values were excluded from the analyses. We also excluded all regions which contain common inversion polymorphisms. The method to estimate the distribution of fitness effects of new mutations (DFE) requires all sites to have been sampled for the same number of chromosomes and since some sites were not successfully sampled in all samples we reduced the original data set to 128 chromosomes by randomly sampling the polymorphisms at each site without replacement. To estimate divergence from *D. yakuba* we randomly sampled one *D. melanogaster* chromosome per site.

Coding exon and short intron (≤65 bp) annotations from *D. melanogaster* were retrieved from FlyBase (release 5.50, http://flybase.org/, last accessed March 2013). Genes 1:1 orthologs across *D. yakuba – D. melanogaster* were obtained from FlyBase (http://flybase.org/). We obtained a multiple genome alignment between the DGRP isogenic lines (Mackay, et al. 2012) and the *D. yakuba* genome (Drosophila 12 Genomes, et al. 2007) using the BDGP 5 coordinates. This alignment is publicly available at http://popdrowser.uab.cat/ (Ramia, et al. 2012). For each gene, we took all non-overlapping coding exons, independently of their inclusion levels. When two exons overlapped, the largest was chosen for subsequent analyses. Only exons without frameshifts, gaps or early stop codons were retained. In this way, we tried to avoid potential alignment errors which would inflate our mutation rate estimates. Our final data set fulfilling all of these criteria contains 7,918 coding genes.

Exonic sequences were trimmed in order to contain only full codons. We define our sites “physically”, so we estimated the rates of substitution at sites of different degeneracy separately. Only zero-fold and 4-fold degenerate sites in exon core codons (as described by Warnecke and Hurst (Warnecke and Hurst 2007)) were used. To estimate the rate of synonymous substitutions, we restricted our analysis to those triplets coding the same amino acid in the two species (*D. melanogaster – D. yakuba*). In restricting our analysis to codons not exhibiting nonsynonymous differences we assume that the codon has undergone no amino acid substitution — this avoids having to compute the different pathways between two codons, which differ by more than one change and it is a reasonable assumption given the level of amino acid divergence. For 4-fold degenerate sites we used the method of Tamura (Tamura 1992) to correct for multiple hits; this method allows for unequal GC content and ts/tv bias. We calculated the number of substitutions and the folded site frequency spectrum (SFS) for 4-fold degenerate sites and zero-fold degenerate sites, using a custom Perl Script.

We used positions 8–30 of introns ≤65 bp in length as an alternative neutral reference for some analyses (Halligan and Keightley 2006). For intron sequences, the invariant GT and AG dinucleotides at the 5' and 3' splice junctions, respectively, are excluded before calculating divergence. Only genes with at least one short intron and with less than 10% of gaps in the aligned sequences were kept. 4,676 orthologous genes pass the intron quality criteria in our final data set. We use an *ad hoc* Perl Script to estimate the number of short intron substitutions and to compute the folded SFS. Multiple hits were corrected using the Jukes and Cantor method (Jukes and Cantor 1969).

Details of the assembly and data filtering of the Rwandan DPGP dataset can be found in Campos et al. (Campos, et al. 2012; Campos, et al. 2014). The dataset comprises 22 genomes sequenced from haploid eggs. They excluded regions that showed admixture with European populations (Pool, et al. 2012) but did not exclude tracts of identity-by-descent (IBD). The IBD tracts potentially affected two strains RG10 and RG15 and a total 30MB ((Pool, et al. 2012) supplementary Table 4); it is unlikely that such a small amount of IBD will affect our results. The numbers of synonymous and non-synonymous sites and polymorphisms and the SFS were estimated by Campos et al. (Campos, et al. 2014).

Recombination rates were taken from Comeron et al. (Comeron, et al. 2012). They estimated the rate of crossovers in 100 kb non-overlapping windows in cM/Mb units. The rate of crossing-over used for a gene is that of the rate in the 100kb window that overlapped the mid-point of the gene.

Hypergeometric sampling

To investigate the correlation between p_N/p_S and p_S ranking genes by its recombination rate we split p_S into 2 independent variables (similar to the splitting done in Smith and Eyre-Walker (Smith and Eyre-Walker 2003); Piganeau and Eyre-Walker (Piganeau and Eyre-Walker 2009)). This was done by generating a random multivariate hypergeometric variable as follows:

P_{s1} = Hypergeometric(P_s, ½ x L_s), (1)

$P_{s2} = P_s - P_{s1}$, (2)

where L_s is the total number of synonymous 4-fold degenerate sites and P_s is the total number of 4-fold segregating sites. We divide ΣP_{s1} and ΣP_{s2} by ½ of ΣL_s to get p_{S1} and p_{S2} at the bin level, respectively. We used p_{S1} to estimate p_N/p_{S1} and p_{S2} as a statistically independent estimate of the level of neutral diversity at the bin level.

To investigate the correlation between p_N/p_S and p_S ranking genes by its level of neutral diversity we split p_S into 3 independent variables following the scheme described by James et al. (James, et al. 2017). This was done by generating a random multivariate hypergeometric variable as follows:

P_{s1} = Hypergeometric(P_s, ⅓ x L_s), (3)

$P_{s2\text{-}3} = P_s - P_{s1}$, (4)

P_{s2} = Hypergeometric($P_{s2\text{-}3}$, ⅓ x L_s), (5)

$P_{s3} = P_s - P_{s1} - P_{s2}$ (6)

We divided ΣP_{s2} and ΣP_{s3} by ⅓ of ΣL_s to get p_{S2} and p_{S3} at the bin level, respectively. We used p_{S1} to rank genes and assign genes to bins, we then used p_{S2} to estimate p_N/p_{S1} and p_{S3} as a statistically independent estimate of the level of neutral diversity at the bin level.

Gene binning strategy

To estimate the ratio between non-synonymous (p_N) to synonymous polymorphisms (p_{S1}) (or short intron, p_{i1}) it is necessary to combine data from several genes because estimates from a single gene are noisy and often result in undefined values due to a lack of segregating synonymous (or short intron) sites. We therefore group genes into bins according to their rate of recombination or neutral polymorphism. Three different grouping schemes were used: 10 bins, 20 bins and 50 bins.

To investigate the relationship between the log of p_N/p_{S1} (or p_N/p_{I1}) and the log of our proxy for the $N_e \sim p_{S3}/d_S$ (or p_{I3}/d_I) we used the substitution rate at exon core 4-fold degenerate sites (or at short intron sites) as a proxy for the neutral mutation rate.

Distribution of fitness effects estimated using the SFS

DFE-alpha (Eyre-Walker and Keightley 2009) models the DFE at functional sites by assuming a gamma distribution, specified by the mean strength of selection, $\gamma = -N_e s$, and a shape parameter β, allowing the distribution to take on a variety of shapes ranging from leptokurtic to platykurtic. We ran DFE-alpha assuming a single instantaneous change in population size from an ancestral size N_1 to a present-day size N_2 having occurred t_2 generations ago using the folded SFS mode since the results are more robust to both demographic changes and linked selection (Messer and Petrov 2013). Provided the SFS at both neutral and functional sites, DFE-alpha infers γ, β, N_2/N_1 and t_2. We ran DFE-alpha (version 2.16) for each bootstrap replicate independently (see below) using the local version provided at: http://www.homepages.ed.ac.uk/pkeightl//software.

Confidence intervals and p-value estimates

To calculate the 95% confidence intervals (CIs) for the slope (b) of the relationship between $\log(p_N/p_S)$ vs $\log(p_S)$, we bootstrapped the data by gene 1,000 times. We split each bootstrap dataset into N bins (see Gene Binning Strategy above) and reestimated the slope for each replicate independently. To estimate the statistical significance of the difference between the slope (b) and the shape parameter (β) of the DFE using the SFS (Eyre-Walker and Keightley 2009) we estimated β and b for each bootstrap dataset. The p-value was the proportion of replicates in which $-b > \beta$.

Statistical analyses

All statistical analyses are performed using the R statistical software (Team 2013). Linear regressions are carried out calling the R function “lm” (from the R package “base”). We calculate Spearman rank correlations (ρ) using the R function “cor.test” (from the R package “base”). The random hypergeometric

variable is obtained through the R function “rhyper” (from the R package “stats”). All the code used to perform the analyses are available upon request from DC.

Forward simulations

We used the forward simulation software SLiM, version 2.4.1 (Haller and Messer 2017) to simulate the behaviour of a single, 20 kb non-recombining locus with a mutation rate of 1×10^{-6}, evolving over time in a small population of 100 individuals. Parameter values were chosen to generate substantial variation in the locus.

We ran two types of simulations. In the first, we investigated the impact of selective sweeps on different types of site, from strongly deleterious to neutral. Our locus was composed of 9 equally common categories of site at which mutations were subject to negative selection with $4Ns$ values of 0, -1, -2, -4, -8, -16, -32, -64, -128. We also allowed the locus to undergo adaptive evolution varying the total proportion of adaptive mutations from 0 to 0.5% We then investigated how the genetic diversity at each category of site changed as we increased the frequency of selective sweeps, such that adaptive mutations constituted between 0.05% and 0.5% of the total number of mutations that occurred in a simulation run. The adaptive mutations were subject to selection such that $4Ns = 400$. For each simulation run, after an initial burn-in of 1000 generations we took samples of all mutations segregating in the population every 500 generations, corresponding to $5N$ generations so that samples should be largely independent. The simulations, were run for 25500 generations in total. For each set of parameters, we ran our simulations 30 times. We then averaged values of diversity over simulation runs. To compare the impact of hitch-hiking on sites under different levels of selection we divide the diversity for a particular set of sites (e.g. $Ns = -4$) by the mean diversity from the simulation with no adaptive evolution.

In the second set of simulations, the locus was divided into two, with half of the mutations being neutral and the other half deleterious mutations which

were sampled from a gamma distribution with a shape parameter 0.4 and a mean *s* of 10, which is roughly in line with what we observe in estimates of the DFE from real data. As before, we altered the frequency at which a locus experienced selective sweeps by changing the frequency with which adaptive mutations occurred. We used 12 different frequencies of adaptive mutations, such that adaptive mutations constituted between 0 and 1% of the total mutations that occurred in a simulation run. We used the same burn-in and sampling procedure as in our previous set of simulations, and again for each set of parameters we ran our simulations 30 times. We then calculated average values of p_S and p_N/p_S across our simulation runs. Code to run and analyse the forward simulations can be found at https://figshare.com/articles/Code_for_producing_the_simulations_in_Nearly_Neutral_Evolution_Across_the_Drosophila_melanogaster_Genome_/3084202 .

Acknowledgements

We are grateful to Meg Woolfit for early work on this project many years ago and for Matt Webster for reinvigorating our interest in this problem. Funding for this study was provided by the University of Sussex and NERC, grant number NE/L502042/1.

Literature cited

Akashi H. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221-238.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139:1067-1076.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149-1152.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A 104:12410-12415.

Begun DJ, Aquadro CF. 1992. Levels of naturally occuring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519-520.
Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. USA 99:13616-13620.
Brandvain Y, Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World. Trends Genet 32:201-210.
Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining regions of the Drosophila melanogaster genome. Genome Biol Evol 4:278-288.
Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. Mol Biol Evol 31:1010-1028.
Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics 22:231-238.
Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.
Chen J, Glemin S, Lascoux M. 2017. Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. Mol Biol Evol 34:1417-1428.
Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in Drosophila melanogaster. PLoS Genet 8:e1002905.
Corbett-Detig RB, Hartl DL. 2012. Population genomics of inversion polymorphisms in Drosophila melanogaster. PLoS Genet 8:e1003056.
Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet 47:126-131.
Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223-227.

Drosophila 12 Genomes C, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203-218.
Ellegren H, Galtier N. 2016. Determinants of genetic diversity. Nat Rev Genet 17:422-433.
Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G. 2010. Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. Genome Res 20:1558-1573.
Eyre-Walker A. 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences of the United States of America 107 Suppl 1:1752-1756.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26:2097-2108.
Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet 12:e1005774.
Gordo I, Dionisio F. 2005. Nonequilibrium model for estimating parameters of deleterious mutations. Phys Rev E Stat Nonlin Soft Matter Phys 71:031907.
Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. Genetics 189:1389-1402.
Haller BC, Messer PW. 2017. SLiM 2: Flexible, Interactive Forward Genetic Simulations. Mol Biol Evol 34:230-240.
Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. Genome Res 16:875-884.
Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. Genetics 169:533-538.
James J, Castellano D, Eyre-Walker A. 2017. DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA. Heredity (Edinb) 118:88-95.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro NH, editor. Mammalian protein metabolism. New York: Academic Press. p. 121-123.
Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymrophism frequencies. Genetics 177:1-11.
Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. Trends Ecol Evol 29:33-41.
Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012. Genomic Variation in Natural Populations of Drosophila melanogaster. Genetics 192:533-+.
Lynch M. 2010. Evolution of the mutation rate. Trends Genet 26:345-352.
Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17:704-714.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401-1404.
Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu DH, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173-178.
Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci U S A 110:8615-8620.
Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93:2873-2878.
Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, et al. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358:951-954.
Ohta T. 1993. Amino acid substitution at the *Adh* locus of *Drosophila* is facilitated by small population size. Proc. Natl. Acad. Sci. USA 90:4548-4551.
Ohta T. 1972a. Evolutionary rate of cistrons and DNA divergence. J. Mol. Evol. 1:150-157.

Ohta T. 1977. Extension of the neutral mutation drift hypothesis. In: Kimura M, editor. Molecular Evolution and Polymorphism. Mishima: National Institute of Genetics. p. 148-167.
Ohta T. 1992. The nearly neutral theory of molecular evolution. Ann. Rev. Ecol. Syst. 23:263-286.
Ohta T. 1972b. Population size and rate of evolution. J. Mol. Evol. 1:305-314.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96-98.
Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. 40:56-63.
Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. 1:18-25.
Pennings PS, Kryazhimskiy S, Wakeley J. 2014. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet 10:e1004000.
Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. PLoS ONE 4:e4396.
Pool JE. 2015. The Mosaic Ancestry of the Drosophila Genetic Reference Panel and the D. melanogaster Reference Genome Reveals a Network of Epistatic Fitness Interactions. Mol Biol Evol 32:3236-3251.
Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, et al. 2012. Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet 8:e1003080.
Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proc Natl Acad Sci U S A 104:13390-13395.
Presgraves DC. 2005. Recombination enhances protein adaptation in Drosophila melanogaster. Current biology : CB 15:1651-1656.
Ramia M, Librado P, Casillas S, Rozas J, Barbadilla A. 2012. PopDrowser: the Population Drosophila Browser. Bioinformatics 28:595-596.
Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in *Drosophila* are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5:704-716.

Smith NG, Eyre-Walker A. 2003. Partitioning the variation in mammalian substitution rates. Mol Biol Evol 20:10-17.
Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. 9:678-687.
Team RC. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. Mol Biol Evol 24:2755-2762.
Watterson GA. 1975. On the number of segregating sites. Theor. Popul. Biol. 7:256-276.
Welch JJ, Eyre-Walker A, Waxman D. 2008. Divergence and Polymorphism Under the Nearly Neutral Theory of Molecular Evolution. J Mol Evol 67:418-426.
Woolfit M, Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol Biol Evol 20:1545-1555.
Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. Proc Biol Sci 272:2277-2282.